Genetic Epidemiology 4

Shaking the tree: mapping complex disease genes with linkage disequilibrium

Lyle J Palmer, Lon R Cardon

Much effort and expense are being spent internationally to detect genetic polymorphisms contributing to susceptibility to complex human disease. Concomitantly, the technology for detecting and genotyping single nucleotide polymorphisms (SNPs) has undergone rapid development, yielding extensive catalogues of these polymorphisms across the genome. Population-based maps of the correlations amongst SNPs (linkage disequilibrium) are now being developed to accelerate the discovery of genes for complex human diseases. These genomic advances coincide with an increasing recognition of the importance of very large sample sizes for studying genetic effects. Together, these new genetic and epidemiological data hold renewed promise for the identification of susceptibility genes for complex traits. We review the state of knowledge about the structure of the human genome as related to SNPs and linkage disequilibrium, discuss the potential applications of this knowledge to mapping complex disease genes, and consider the issues facing whole genome association scanning using SNPs.

Lancet 2005; 366: 1223–34

This is the fourth in a Series of seven papers on genetic epidemiology.

Western Australian Institute for Medical Research and University of Western Australia Centre for Medical Research, University of Western Australia, QE-II Medical Centre, B Block, Hospital Avenue, Nedlands WA 6009,Australia (Prof L J Palmer PhD); and Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK (L R Cardon PhD)

Correspondence to:
Prof Lyle Palmer
lyle.palmer@cyllene.uwa.edu.au

Genomic approaches to disease association mapping

Genomics is transforming epidemiology, medicine, and drug discovery,[1–7] and attention is being directed towards population-based genetic association studies for complex phenotypes.[3,8–12] For many complex conditions, the genetic basis of susceptibility to disease, disease progression and severity, and response to therapy has been increasingly emphasised in medical research, with the ultimate goal of improving prevention, diagnosis, and treatment.[4,5,13,14]

Completion of the human genome sequencing project has been followed by three advances that provide novel opportunities for understanding the pathogenesis of common diseases:[1,15] (1) compilation of extensive catalogues of DNA sequence variants across the human genome (polymorphic loci);[15–17] (2) more rapid and cheaper molecular genetic techniques for investigating polymorphic sites; and (3) increasing availability of large, population-based samples such as the European Prospective Investigation into Cancer and Nutrition,[18] the International Study of Infarct Survival,[19] and the Million Women Study.[20] Large, national cohorts (eg, the UK Medical Research Council and Wellcome Trust Biobank[21]) are attracting funding bodies in many countries. Although the genomics revolution and the generation of high-density single nucleotide polymorphism (SNP) maps has benefited the investigations of mendelian (single-gene) diseases, our discussion will be restricted to common complex conditions such as obesity and cardiovascular disease that are determined by multiple genetic and environmental factors. Such diseases constitute the main health burden in developed countries.[1,4,5,22]

Given the rapidly changing nature of the field of genetic epidemiology, the large amounts of genomic data being generated at considerable cost, as well as the apparent and unforeseen obstacles facing progress, it is important to consider these intiatives in the context of expediting the discovery of complex human disease genes. We review knowledge about the human genome as related to SNPs and linkage disequilibrium (LD), discuss the potential applications of this knowledge to mapping complex disease genes, and look at the feasibility of whole genome association using SNPs.

Genomic information in mapping complex disease genes

We are at the beginning of our ability to map complex disease genes. Sequencing of the human genome remains the key to this enterprise, but the focus of that project was the consensus human sequence, which by definition cannot contain information about individual differences of medical relevance.[23] To make use of the consensus sequence, the SNP Consortium was formed in 1999, with other public and private projects, with the aim of discovering common polymorphism sites in the human genome.[24] The increasingly complete catalogue of common genetic variants that is being applied to association studies of complex phenotypes is a direct extension of the consortium's work. The natural next step to the SNP discovery phase was to genotype identified SNPs in individuals to begin to assess their potential usefulness for disease mapping. This work is ongoing in the International HapMap project. The next stage will involve applications to gene discovery. Some genes associated with complex diseases have been discovered by association-based genetic mapping.[25,26] Genetic association studies are discussed in detail in other papers in this series.[27,28]

The association of an allele with a phenotype due to correlation (ie, LD) between the allele and a nearby causal variant—so-called indirect association—is the main thrust of whole-genome association studies and large-scale genomic projects like the International HapMap project (discussed below).

SNPs

Because the mutation rate is low (around 10^{-8} per site per generation) when set beside the most recent common ancestor of any two people (around 10^4 generations),[25] most SNPs are thought to arise from a single historical mutational event. Across the human genome, there are far more SNPs than any other types of polymorphism[29]—at least 10 million SNPs with frequency greater than 1%, yielding an average spacing of one every 290 base-pairs.[30] These common SNPs are thought to account for around 90% of human genetic variation.[17,31–33]

There are four important advantages of using SNPs rather than other types of genetic polymorphism to investigate the genetic determinants of complex human diseases.[8,34,35] First, SNPs are plentiful throughout the genome, being found in exons, introns, promoters, enhancers, and intergenic regions,[36,37] and some of these polymorphisms might themselves be functional. Second, groups of adjacent SNPs might exhibit patterns of correlations that could be used to enhance gene mapping[38] and which may highlight recombination hot-spots.[39] Third, interpopulation differences in SNP frequencies can be used in population-based genetic studies.[40,41] Fourth, SNPs are less mutable than other types of polymorphism,[42,43] and this greater stability could allow more consistent estimates of gene-phenotype associations.

The common SNPs have been subject to large cataloguing projects funded by both government and industry.[16,17,44] These efforts have involved targeted SNP discovery by mutation detection[45] or primary resequencing in candidate genes or regions.[30,31] Of more than 10 million SNPs so far identified, more than 5 million have been validated.

See http://www.ncbi.nlm.nih.gov/ SNP/index.html

Many other SNPs present in major ethnic groups are likely to be discovered. SNP databases are constantly being updated (panel 1).[17,31] However, the data are not infallible, as some putative polymorphisms turn out to be sequencing errors or rare or population-specific variants often not detected in subsequent studies.[16,17] Limitations due to cost and the incomplete status of SNP databases mean that the association analysis of SNPs in complex disease genetics has been mostly limited to polymorphisms within biologically plausible candidate loci. Many investigators interested in specific genes or pathways have independently sought to identify sequence variants by primary resequencing in their own study populations.[31,46]

SNPs are finding widespread use in fine mapping of genetic disorders, in the delineation of genetic influences in multifactorial diseases such as breast cancer, cardiovascular disease, type 2 diabetes and asthma, and as genetic markers to predict responses to drugs and adverse drug reactions.[22] There are at least six primary areas of potential application for SNP technologies in improving our understanding of complex disease: (1) hypothesis-free gene discovery and mapping; (2) association-based candidate polymorphism testing; (3) pharmacogenetics; (4) diagnostics and risk profiling; (5) prediction of response to non-pharmacological environmental stimuli; and (6) homogeneity testing and epidemiological study design.[9] There are thus dual imperatives to develop advanced technologies to detect and genotype SNPs, and for improved statistical approaches and study designs to enable SNP data to be incorporated into epidemiology and clinical medicine.

***Panel 1:* Selected websites**

SNP databases

- dbSNP Polymorphism Repository [http://www.ncbi.nlm.nih.gov/SNP/].
- Cancer Genome Anatomy project [http://cgap.nci.nih.gov/].
- Génome Québec [http://www.genomequebec.com/index_e.asp].
- The Golden Path [http://genome.ucsc.edu].
- Human Genome Variation Database [http://hgvbase.cgb.ki.se/]
- The Human Genome Variation Society [http://www.genomic.unimelb.edu.au/mdi/].
- Human Gene Mutation Database [http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html]
- Human SNP Database [http://www-genome.wi.mit.edu/SNP/human/index.html].
- The International HapMap Project [http://www.hapmap.org/]
- LocusLink [http://www.ncbi.nih.gov/LocusLink/]
- NHLBI Programs for Genomic Applications Resources [http://pga.lbl.gov/PGA/PGA_inventory.html]
- OMIM: Online Mendelian Inheritance in Man [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM]
- SNP Consortium [http://snp.cshl.org/].
- SNP View [http://snp.gnf.org/].
- The Sanger Centre [http://www.sanger.ac.uk/].

Software

- An extensive list of genetic analysis software http://linkage.rockefeller.edu/soft/list.html].

Linkage disequilibrium

Most SNPs lie outside genes and are not likely to alter gene structure or function, so they might not be directly associated with any change in phenotype.[47] We need to know whether the DNA sequence variant under consideration is potentially directly functional (ie, could lead to the observed biology) or is indirectly correlated with another DNA sequence variant that is the actual cause of the phenotype of interest. Since candidate genes are usually difficult to select[12] and since functional data are rarely available for a given SNP, testing for indirect association is the model which most attempts at gene

discovery use. LD is discussed in other papers in this series.[27,28] Loci in LD are generally close together, but the relation varies (figure 1). When a variant is first introduced into a population by mutation, it will be perfectly correlated with nearby variants, but over successive generations meiotic recombinations will break up the correlations, and LD will decay (figure 1). Indirect association mapping relies on LD in the sense that the functional variant need not be studied at all, so long as one measures a variant that is in LD with it.

Many factors can influence LD, including genetic drift, population growth, admixture, population structure, natural selection, variable recombination and mutation rates, and gene conversion.[49,50] The International HapMap project was started to describe disequilibrium patterns in some ethnic groups and it should help clarify the value of SNPs for the indirect association mapping of disease genes[25] (see below).

Haplotypes and haplotype estimation

Indirect association mapping by LD relies on gene-phenotype associations at the level of population,[51] and requires a dense map of markers.[52] It may be enhanced by examining multiple markers simultaneously or using haplotypes, which are linear arrangements of closely linked alleles on the same chromosome inherited as a unit. Haplotype analysis in the context of disease association studies is difficult,[53] but haplotypes do contain at least as much information as the genotypes at each component locus, so may prove essential for some disease gene studies.

For M biallelic markers there are 2^M possible haplotypes (though often many fewer are evident), and because we usually do not know in advance which haplotypes might be associated with disease, all are tested. Testing SNPs one at a time would require M tests so the greater information in haplotypes is offset by the cost of testing more of them. The growing use of phylogenetic approaches derived from population genetics in human gene discovery investigations holds promise in this area,[54] as it helps to form natural groupings of haplotypes.

When LD is high, the redundancy amongst markers means that haplotypes can be used in association studies to efficiently map common alleles that might influence the susceptibility to common diseases, as well as for reconstructing genomic evolution.[55] When LD is low, haplotypes will generally be useful in refining SNP-phenotype associations only if they help delineate rare allele frequencies or if there are significant interactions among the SNPs in their effect on the trait. In complex diseases, where multiple variant loci contribute to disease susceptibility, haplotypes are therefore also potentially important since different combinations of particular alleles in the same gene may act as a meta-allele or meta-SNP and have different effects on the protein product and on transcriptional regulation.[56]

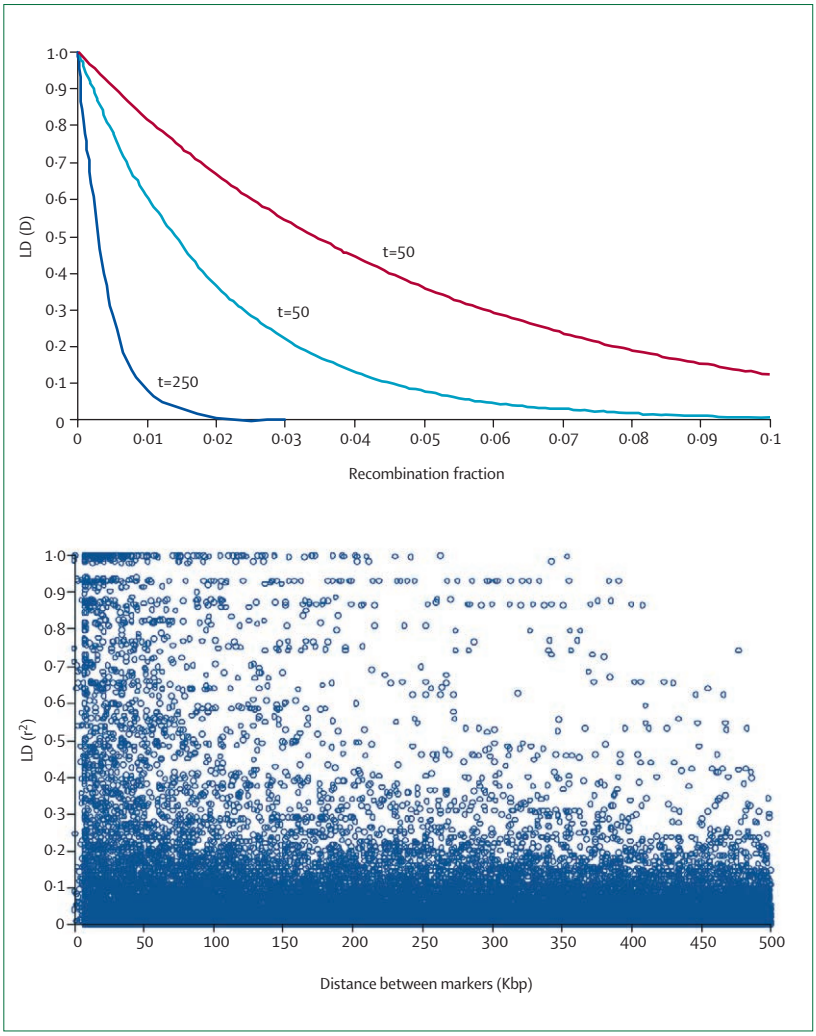

Figure 1: **Theoretical (upper) and observed (lower) patterns of LD decay**
Data from chromosome-wide study of human chromosome 22.[48] Upper: hypothetical decay in LD as function of recombination fraction between two loci. Three curves indicate different time-scales (numbers of generations) since initial mutations that generated markers. For two markers that recombine at rate θ, correlation between two markers is reduced (by 1–θ) in each generation, so at generation t, remaining disequilibrium, $E(D_t)=(1-\theta)^t$. Lower: decay trends in real data from chromosome 22. General shape of theoretical decay apparent in empirical data, but there is a vast amount of variability so that knowing average decay gives little information about any specific pair of genetic loci.

In population-based studies based on unrelated individuals, the parental origin of each allele of a genotype is not known (so-called phase unknown status); haplotypes for double heterozygotes are uncertain and must be estimated.[57] Statistical methods and software are available to estimate haplotypes from phase unknown genotype data in large population-based samples of unrelated individuals or in family data,[57–62] and new maximum-likelihood methods have been developed to allow the testing of statistical association between haplotypes and binary, ordinal, and quantitative traits.[63] However, the use of haplotypes derived from phase-unknown genotype data is not always straightforward, and the value of these techniques for gene mapping is not yet clear.[57,58,65]

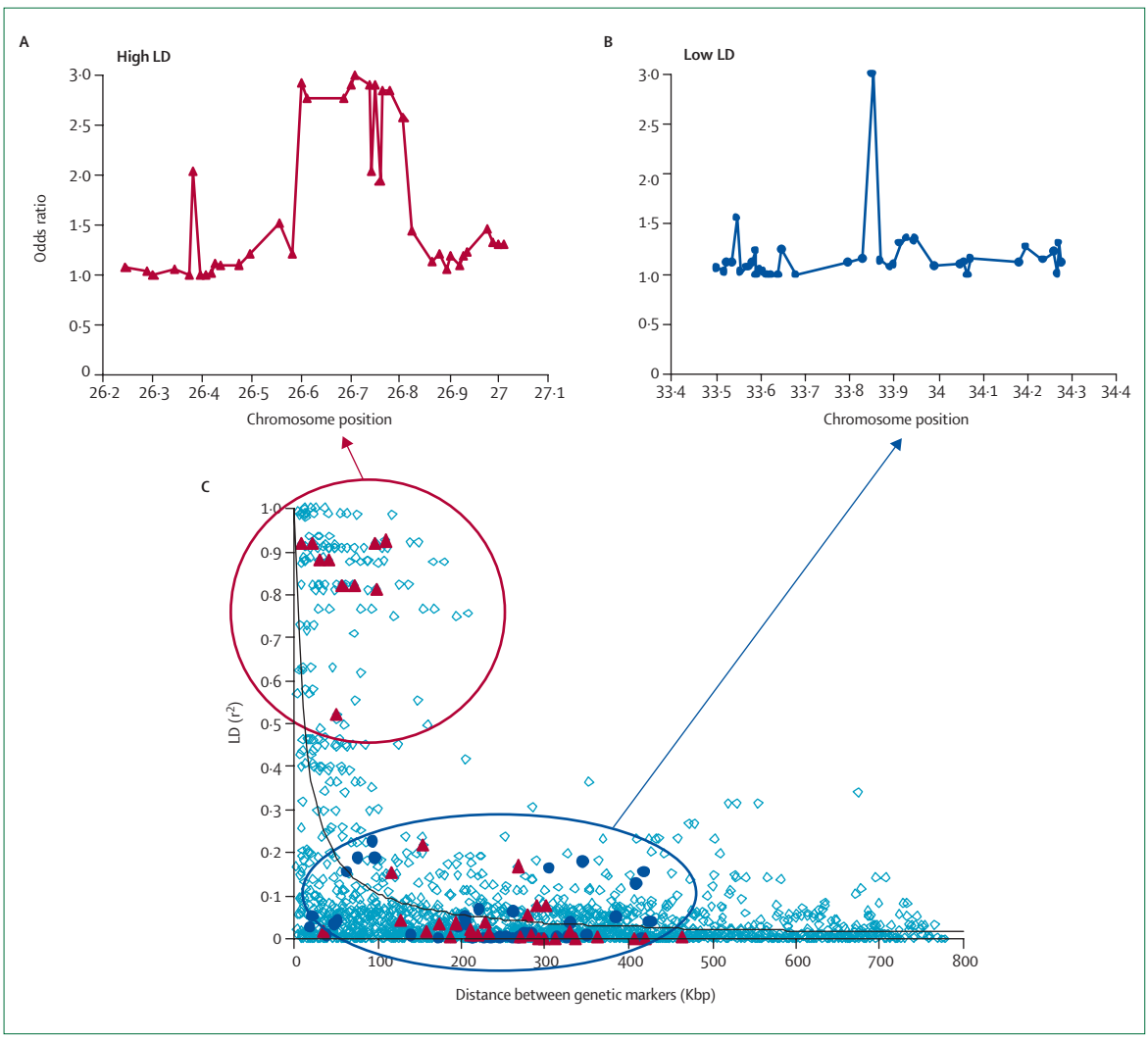

Figure 2: **The role of LD in facilitating allelic association**
A and B: disease association profile for hypothetical disease in which aetiological locus confers OR of 3·0. Markers in A show extensive background LD, so many are associated with trait. Markers in B show little LD, so only causal locus is associated. Distribution of LD for these two scenarios shown below to illustrate that knowing local patterns can help to delineate expected patterns of association and design efficient novel studies. Data from chromosome 22, in which arbitrary locus was designated disease gene in high and low regions of the chromosome.[48] Decay in odds ratio computed as described.[26]

LD patterns across the genome

Large sets of SNPs and improved genotyping technology and statistical methods for haplotype estimation are necessary for improving gene discovery via indirect association analysis, but there is more information available. The extreme variability in the correlation between physical distance and LD in a given genomic region (figure 1) means that two genetic variants that are physically close will sometimes be completely independent, whereas loci that are very far apart will sometimes be highly correlated. Thus, when LD is low, screening nearly all of the SNPs in a given region could still miss the relevant locus. When LD is high, evidence for association can be found for most of the loci examined, which would reveal little about the precise localisation of the aetiological variant. These two extremes are depicted in figure 2, where a chromosome region in which many markers are associated with the outcome (top left) is contrasted with a region in which only a single marker reveals evidence for association. The different patterns of disease association are due to different LD patterns in the chromosome regions.

Until recently, little was known about LD patterns in the genome except for a few well-characterised genes and gene families. However, studies of large genomic regions or entire chromosomes are now adding to this knowledge base, highlighting the importance of dense marker panels and revealing extensive variability in LD patterns and recombination rates.[66–70] Further information is needed to enable appropriate study design and more accurate interpretation of association studies. The International HapMap was initiated in recognition of this need (panel 2).

Panel 2: the International HapMap project

This large project aims to construct genome-wide maps of LD patterns in multiple populations.[71] The project calls for genotyping up to a density of more than 1 SNP per 1000 bp in samples collected in the USA, Nigeria, China, and Japan. In validating such a broad spectrum of SNPs and in building these high density maps, the HapMap project aims to facilitate genetic mapping across a broad array of complex phenotypes, including those relevant to diagnostic and therapeutic applications. Importantly, the raw data are being released publicly, allowing immediate use of the emerging maps by the scientific community. The project will also foster development and application of different statistical methods for LD mapping.

The main practical objective of the HapMap project is to identify sets of SNPs that will take advantage of the LD patterns identified to allow more efficient genotyping.[71] When LD is high, the redundancy between markers implies that most of the information can be captured without genotyping all markers. Non-redundant markers that capture most or all of the LD information in a given genomic region have been termed haplotype tagging SNPs.[72,73] Defining and genotyping a relatively small number of these SNPs could allow unambiguous determination of the common haplotypes in a population, and capture all or most of the LD within that region.[72,74–76,77,78,79,80] By this means, SNP-phenotype association studies can be done relatively efficiently, by contrast with genotyping all common variants in a given genomic region or in the entire genome.[72] Various statistical approaches have been developed to define haplotype-tagging SNPs,[73–77,79–83] though it is not yet known precisely how much saving they might yield: estimates for European samples vary widely from about a tenth to a fiftieth of the 10 million common SNPs.[46,71,73,84]

Methodological and study design issues

Increasingly complete SNP databases, better genotyping, high density LD maps, and large, population samples are essential for complex trait association studies but do not guarantee success. Other obstacles remain,[85,86] many of which are outside the investigator's control. Examples, reviewed elsewhere, include technical issues in genotyping, limitations to our understanding of LD,[49,87] and difficulties in investigating gene-phenotype associations involving multiple interacting genetic and environmental factors.[12,35,88] However, in this section we will highlight some additional factors emerging from the ongoing integration of the large-scale genetic and epidemiological data.

Statistical methods

The focus on SNP genotyping has made it clear that new statistical methods are needed for LD mapping of complex trait genes,[12,85,89,90] and has led to re-examination of mapping methodologies and study designs.[10,12,49,52,91] The fundamental issue of how to deal with the volume of data produced is only now being addressed; developments in biostatistics have been lagging behind the capacity to generate SNP genotypes.[74,92] The best way to apply SNPs and LD mapping data to the genetic epidemiology of common diseases remains unclear. A number of statistical methods for selecting haplotype-tagging SNPs are available and more are in the pipeline.[75,76,93] The differences between these diverse approaches will need to be understood to make efficient use of genome-wide LD data. Additionally, the applicability of a tagging approach developed in one population to other populations has not yet been fully examined, leading in part to the wide range of differences in the estimates of the potential gain in genotyping efficiency resulting from the use of htSNPs.

One practical challenge facing haplotype tagging (panel 2) is the definition of the genomic region to be tagged. Tagging was initially described as a means of efficiently genotyping,[72] but it was later wedded to the notion of haplotype blocks, which are regions of very high LD delineated by regions of low LD.[74,94,95] As block boundaries are not always consistent within or between populations[67,77,96] or between statistical definitions it is not clear that block-tags defined in one sample will capture the same information in another. Ultimately, the region definition problem may be addressed empirically by examining multiple samples drawn from many populations, or theoretically by statistical methods that do not depend on physical boundaries.[81]

Missing data, an issue for genetic analysis generally, are a particular problem for haplotype analysis. Sequencing or genotyping a given set of SNPs is rarely 100% complete and missing data with each additional SNP included in a haplotypic analysis. Other branches of statistical investigation have learned that ignoring missing data or restricting analysis to individuals with complete data can lead to biased or inefficient analyses, even when data are missing completely at random.[97–101] This problem worsens if data are not missing at random, as may be the case with systematic errors in genotyping assays. Methods for dealing with missing data have seldom been applied to genetic epidemiology but more needs to be known about the extent to which missing data are a problem in genetic association analyses of SNPs and haplotypes and about the application of methods for dealing with missing data in such studies.

Power, p values, and multiple testing

In complex disease genetics, both type I and type II error needs to be reduced.[8,12,102] Power for studies of allelic association will depend primarily upon sample size, the effect size of the susceptibility locus, the strength of LD with a marker, and the frequencies of susceptibility and marker alleles.[26] Figure 3 illustrates sample sizes needed to detect a true odds ratio of 1·5 with 80% power and

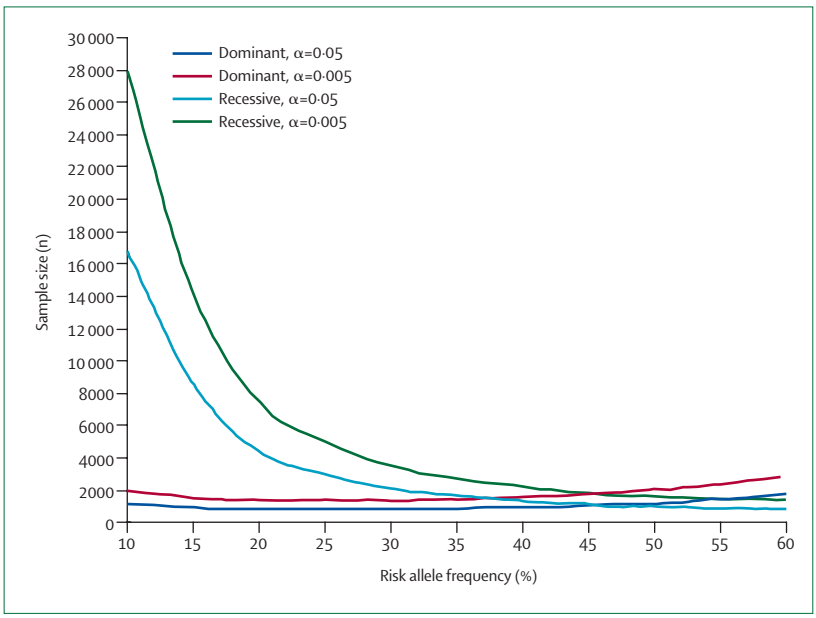

Figure 3: **Sample size estimates for case-control analyses of SNPs**
Sample size is cases plus controls, with one control per case; detectable difference of OR >1·5; power 80%. α=type I error rate.

type I error probability (α) of either 0·05 or 0·005. Even for the best-case scenario, a common SNP acting in a dominant fashion, more than 800 people are needed at the 0·05 level, which is still in widespread use by researchers and journal editors.

Multiple testing is an issue in many genetic association studies of candidate loci where multiple SNPs in one gene or multiple SNPs in several loci are tested, or both,[103] and an α on the order of 0·005 could be more realistic, even for only a small set of genetic markers. Use of α=0·005 or with an uncommon SNP that acts in a recessive fashion leads to large sample sizes. This problem will be exacerbated in studies with more SNPs, such as whole genome association designs (and even smaller values of α) where the numbers become higher still. These power calculations show that the sample sizes used in many case-control association studies of complex phenotypes have been too small to detect even quite a large effect of an SNP. Genetic association studies have generally been underpowered,[4,10,52,85,104] and future studies will have to be much larger for most human diseases. Very large cohort studies will be needed for genetic epidemiological investigations of many common conditions,[11,21] and collaboration among research groups is becoming increasingly important.

The testing of large numbers of SNPs for association with one or more traits raises important statistical issues about false-positive rates and levels of statistical significance.[52] Post-hoc corrections tend to be too conservative, especially since many (such as a simple Bonferroni correction[105]) do not take proper account of the correlation between SNPs in LD with each other. Haplotype-tagged SNPs are chosen to be as independent as possible. So, use of tag SNPs requires more stringent correction than studies of equal numbers of correlated genomic SNPs. With correlated markers, statistical techniques to correct for multiple comparisons are emerging,[106] but replication of genetic association findings in independent population samples remains the gold standard for complex disease genetics.[8,107,108]

Population heterogeneity

Population heterogeneity is a serious issue for gene discovery in any population-based study of complex diseases.[109–114] Disease prevalence often changes with geography and ethnic origin, and allele frequencies can vary widely throughout the world.[115] Additionally, there is likely to be a high degree of variation in LD between populations of different origins,[112,116] and between different genomic regions,[48,68,69,117,118] leading to differences in genetic-physical map correlations, estimates of LD and haplotypes, tagging SNP selections, and other outcomes. This heterogeneity can complicate or even prevent gene discovery and cloud apparent evidence for replication.

For association studies of many complex diseases, case-control designs have become the approach of choice. The biggest criticism of such studies has been the potential for undetected population stratification: spurious association may arise when allele frequencies vary across subpopulations (eg, people from different ethnic groups[119]). This is a potential issue for both direct candidate gene approaches and indirect association.[120] Such stratification may result from recent admixture or from poorly matched cases and controls. Genomic control, genotyping of random panels of SNPs to assess population structure and begin to correct for it,[113,114,121–127] coupled with careful population-based studies of unrelated controls should reduce confounding by population stratification.[128] Research on the performance of genomic control with large samples has revealed that the larger the sample size, the greater the potential bias from stratification.[113,114] We may need to type many hundreds or even thousands of markers to detect and control subtle stratification in large samples.[113] Fortunately, genotyping costs are falling.[113,114]

Understanding how aetiological factors act at a population level will be a critical step for the clinical application of knowledge about the genome.[4,129,130] Genetic knowledge will only become clinically useful when it is placed back in an epidemiological and public health context.[5–7,13,131] Very large, longitudinal, well-characterised population-based studies drawn from multiple ethnic groups will have a vital role in the implementation of SNP-based gene discovery and in diagnostic tests for complex phenotypes in the outbred, highly admixed populations that increasingly characterise human societies today.[73]

Rare alleles

Current attention in population-based association studies is focused almost entirely on genetic markers and aetiological variants that are common (>1% frequency). This is true for SNP detection studies ,[24] public databases,[31,67,96,132] the HapMap project,[71] and haplotype tagging, and most sample collections are powered to detect effects only arising from common variants. There are several reasons for this emphasis. The most cited one relates to the common-disease, common-variant hypothesis, which holds that genetic influences on diseases of high population prevalence are old, and are thus typically very common. There are arguments and evidence for and against this hypothesis, as well as empirical support and counterexamples.[133–135]

Another reason for the emphasis on common alleles is purely practical. Common diseases are assumed to be influenced by many genetic and environmental factors, all with a modest effect on the trait. If the genetic influences are rare, the sample sizes required to detect the modest effects become impossibly large[8,26,136] (figure 3). Thus, in the absence of so-called low-hanging fruit (genes with major effects on complex phenotypes) it is impractical to search for rare genetic effects using the allelic association design. This practical consideration explains the current focus on gene discovery strategies aimed at common alleles and implies that real effects associated with rare alleles will go undetected.

Allelic heterogeneity accentuates the problem of rare alleles. With the breast and ovarian cancer loci *BRCA1* and *BRCA2*, the phenotype results from a very large number of different mutations in the same gene(s),[137] so that many people have extremely rare or unique mutations. Such heterogeneity would possibly not be detected by population-based association, no matter how large the sample size or the number of common SNPs genotyped (the *BRCA1* and *BRCA2* loci were identified by family-based linkage[138,139]) Thus there are genetic aetiologies that are not amenable to discovery by population association analysis.[88,135] As these are not known a priori, it is important to emphasise that the vast SNP datasets being constructed, the HapMap project, enhanced genotyping capacity, and all the other resources being brought to bear on this problem will not always lead to gene discovery.

Replication

Several recent articles have addressed the features of a good genetic association study.[12,26,73,107,140] This focus on study design stemmed from the realisation that genetic association studies of complex phenotypes often either fail to discover susceptibility loci or fail to replicate studies that did.[12,73,85,88,141–143] Despite the widespread use of genetic case-control studies, their inconsistency is a generally recognised limitation.[84,88] This lack of reproducibility is often ascribed to small samples with inadequate statistical power, biological and phenotypic complexity, population-specific linkage disequilibrium, effect-size bias and population stratification.[8,88,144,145] Other reasons for the non-replication of true positive association results include inter-investigator and inter-population heterogeneity in study design, analytical method, phenotype definition, genetic structure, environmental exposures, and markers genotyped. It is now routinely argued that large sample sizes (generally, thousands rather than hundreds), rigorous p-value thresholds, and replication in multiple independent datasets are necessary for reliable results.[4,26,88,140,143] For most complex human diseases, the reality of multiple disease-predisposing genes of modest individual effect, gene-gene interactions, gene-environment interactions, heterogeneity of both genetic and environmental determinants of disease and low statistical power mean that both initial detection and replication will likely remain difficult.[12,52,85]

Ironically, the advances in SNP genotyping and LD mapping that offer promise for association studies also highlight some of the difficulties that large SNP studies face. Decreasing costs mean that more SNPs will be typed, and thus more spurious results will be obtained. This places a greater burden on establishing robustness via replication. However, different definitions of replication are emerging. Descriptions of so-called confirmatory replication are often attached to findings that appear non-confirmatory. For example, different genetic markers are significantly associated in the follow-up study differ from those in the original report; or the same genetic markers are reported in both studies, but with opposite alleles (ie, the disease allele is the protective allele in the follow-up); or different phenotypes as examined in the initial and follow-up studies. The problem with these definitions is that although they might indicate false positives they could indeed reflect genuine replications because there are genetic reasons for them. For the three examples given above, allelic heterogeneity could explain the first scenario, different population backgrounds the second (as apparent in animal models of disease[146,147]), and the third is consistent with genetic pleiotropy, where one gene influences many phenotypes. Standardised definitions of replication is needed because some explanations (eg, a risk allele in one sample appearing as protective in another sample drawn from the same population) look biologically less plausible than other replication scenarios. Although there is no disputing the importance of heterogeneity within and between samples and genes, there is a risk that heterogeneity could be abused to rationalise negative follow-up studies in positive terms.

In general, studies showing similar results in terms of phenotypes tested and specific SNP associations found offer strong evidence for association. However, those lacking such clear overlap, even with positive association evidence, may require validation using other strategies

or datasets. Future studies of large numbers of SNPs will need to approach these issues carefully lest replication lose its status as a gold standard for genetic association.

Whole genome association

High density SNP maps and the identification of genes by the Human Genome Project[148] have made whole genome association analyses technically feasible for many conditions.[149] However, despite costs heading down to US$0·01 per genotype[150] (a target once regarded as highly ambitious), testing all of the 10 million common SNPs would cost at least US$100 000 per individual or US$200 million for a single study of 1000 cases and controls. Exhaustive genotyping for association is therefore currently impractical.

What is a whole genome association study?

Forms of whole genome association are now being explored.[151] Whole genome implies complete coverage but not all such analyses are the same. For example, marker sets of 100 000 or more SNPs are now commercially available as whole genome panels. In constructing such panels, one could select SNPs in a variety of ways—eg, with a focus on genes only,[152] via haplotype tagging or at random throughout the genome. None of these covers all variation in the genome, so by a strict definition, none offers a whole genome study. Indeed, genotyping 100 000 SNPs in many populations would probably cover less than 50% of common genetic variants.[153] Whole genome association studies will require qualifiers describing their aims, assumptions and presumed coverage. The concern is not so much that what they do find will be false but how many and of what composition are the genetic variants that they missed.

Complete resequencing of the entire genomes of case and control individuals would be ideal, but this technology is not yet available or affordable. The high-density panels being genotyped in the International HapMap project (panel 2) and in industry[70] offer the most immediate form of whole genome coverage. Although rare variants are under-represented, 85–90% of the genetic variants that are common in the samples evaluated may soon be available for disease-gene research.

Reducing the genotyping burden

There are at least two strategies for reducing the number of SNPs that need to be genotyped,[37] one based on indirect association and haplotype-tagging SNPs across the genome (map-based) and the other based on direct association and the genotyping of all potentially functional SNPs across the genome (sequence-based).[22,153]

The map-based approach makes no assumptions about the genes involved or the type of the mutation, though it does assume that disease alleles or haplotypes are sufficiently frequent to have been captured by the original tagging study. Estimates for the number of tag SNPs needed to represent most common variants across the entire human genome range from 200 000 to more than a million.[71,73,78,81,154–156] A single genome-wide study would still cost several million US$ for 1000 cases and controls. Moreover, the SNPs genotyped in such a study would be highly selected in order to reflect the underlying LD patterns in the relevant population. In this regard, the feasibility of whole genome association scans in the map-based model depend critically upon knowledge of genomic LD patterns in multiple populations.[78,87] Random sets of uniformly spaced SNPs, though cheaper, easier to genotype and increasingly available commercially, do not yield the same efficiency or robustness.[71] Further decreases in genotyping cost or savings in the number of markers to genotype are needed for well-powered association studies across the genome.

The sequence-based approach makes savings by assuming that specific variants are more likely to influence complex traits than others. Prioritised lists of such variants[8,22] decrease the number of SNPs to 50 000–100 000 and study costs less than US$1–2 million for 1000 cases and 1000 controls. However, despite the availability of over 10 million SNPs in public databases, further work may be needed to identify all SNPs at the top of the priority list (ie, non-synonomous, non-conservative coding changes[8]). In addition, many coding changes are rarer in their allele frequencies than non-coding changes, thus creating sample size challenges unless the genes have large effects.

One approach that can reduce genotyping requirements under both strategies is the use of generic or universal controls—or a large set of representative controls from which subsets are matched to individual disease samples.[128] Genotyping a genome-wide set of markers on such a sample would allow re-use of the genotypes across the disease samples. Genomic control could facilitate matching and reduce potential confounding.[128] One potential role for large cohort initiatives such as UK BioBank will be to provide such universal controls. Another labour-saving strategy is staged genotyping, so that not all markers are genotyped on all individuals. By genotyping all markers on a subset of the sample and liberally selecting the marker set to be genotyped on the remainder of the sample, it should be possible to retain most of the statistical power while reducing the genotyping load.[157] Savings of up to 75% of potential genotyping reactions with minimal loss of power have been demonstrated with genetic analysis of type 1 diabetes samples.[157]

The map-based and sequence-based approaches both hold promise for genome-wide studies. It is not clear which will prove more fruitful, and it is certain that no single approach will work for all situations.

The future

Explosive growth in technical capacity and genomic knowledge has been tempered by initial failures to find genes for complex phenotypes using any strategy and our

statistical methods and informatics capabilities lag far behind our ability to produce huge amounts of genomic data. What have we learned over the past decade of linkage mapping and association analyses? One important lesson is that everything in human genetics is context specific—specific to the population, environmental exposures, genomic region, and gene under investigation. There is no one paradigm for gene discovery and no single ideal study design or analytical approach. Despite ex cathedra statements on optimum study design and analytical, it is clear that flexible, mixed approaches and hypothesis-free designs are desirable. The genomics revolution has been accompanied by an unfortunate tendency to hyperbole. This has led to unrealistic expectations among clinicians and to cynicism and pessimism within the genetics community. For genetics researchers, one of the most important tasks now is to not add to the hyperbole but to establish and communicate realistic expectations.

Where does LD-based association mapping stand today? For most complex human diseases, the reality of multiple disease-predisposing genes of modest individual effect, gene-gene interactions, gene-environment interactions, inter-population heterogeneity of both genetic and environmental determinants of disease, and low statistical power mean that both initial detection and replication are likely to remain difficult.[12,52,85] However, our understanding of the complexity of the task is improving and new tools and a growing knowledge base (eg, rapid progress in SNP detection, complete catalogues of SNPs, and the attention being paid to methodological problems in LD mapping and haplotypic approaches) do offer prospects for success in gene discovery. These and other developments, taken together with a small but growing number of successful gene localisations for complex phenotypes, suggest that cautious optimism about discovery of genes underlying common human diseases is justified. Another cause for hope is the assimilation of genetic epidemiology into mainstream epidemiology and public health in many academic institutions. The involvement of epidemiologists should improve some of the difficulties that have plagued complex disease genetics, many of which can be blamed on poor design and overinterpretation of marginal results. Our understanding of complex disease pathophysiology has already begun to enter into the realm of clinical genetics,[158] and we have every reason to anticipate that the impact of genomics upon clinical practice and upon our understanding of biology and epidemiology will continue to accelerate.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

This work was supported in part by the Wind-over-Water Foundation (LJP) and by a Wellcome Trust Principal Research Fellowship and NIH grant EY-12562 (LRC).

References

1 Khoury MJ. Genetic epidemiology and the future of disease prevention and public health. *Epidemiol Rev* 1997; **19:** 175–80.
2 Nagy A, Perrimon N, Sandmeyer S, Plasterk R. Tailoring the genome: the power of genetic approaches. *Nat Genet* 2003; **33** (suppl)**:** 276–84.
3 Zerhouni E. Medicine. The NIH Roadmap. *Science* 2003; **302:** 63–72.
4 Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. *Nat Rev Genet* 2003; **4:** 937–47.
5 Merikangas KR, Risch N. Genomic priorities and public health. *Science* 2003; **302:** 599–601.
6 Kelada SN, Eaton DL, Wang SS, Rothman NR, Khoury MJ. The role of genetic polymorphisms in environmental health. *Environ Health Perspect* 2003; **111:** 1055–64.
7 Shostak S. Locating gene-environment interaction: at the intersections of genetics and public health. *Soc Sci Med* 2003; **56:** 2327–42.
8 Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; **405:** 847–56.
9 Schork NJ, Fallin D, Lanchbury JS. Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin Genet* 2000; **58:** 250–64.
10 Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361:** 598–604.
11 Wright AF, Carothers AD, Campbell H. Gene-environment interactions: the BioBank UK study. *Pharmacogenomics J* 2002; **2:** 75–82.
12 Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001; **2:** 91–99.
13 Burke W. Genomics as a probe for disease biology. *N Engl J Med* 2003; **349:** 969–74.
14 Johnson JA. Pharmacogenetics: potential for individualized drug therapy through genetics. *Trends Genet* 2003; **19:** 660–66.
15 Venter JC, Levy S, Stockwell T, Remington K, Halpern A. Massive parallelism, randomness and genomic advances. *Nat Genet* 2003; **33** (suppl)**:** 219–27.
16 Varmus H. Genomic empowerment: the importance of public databases. *Nat Genet* 2003; **35** (suppl 1)**:** 3.
17 Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. *Nat Genet* 2003; **33:** 457–58.
18 Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann Oncol* 1992; **3:** 783–91.
19 First International Study of Infarct Survival Collaborative Group. Randomised trial of intravenous atenolol among 16 027 cases of suspected acute myocardial infarction: ISIS-1. *Lancet* 1986; **2:** 57–66.
20 The Million Women Study Collaborative Group. The Million Women Study: design and characteristics of the study population. *Breast Cancer Res* 1999; **1:** 73–80.
21 Austin MA, Harding S, McElroy C. Genebanks: a comparison of eight proposed international genetic databases. *Community Genet* 2003; **6:** 37–45.
22 Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003; **33** (suppl)**:** 228–37.
23 Cardon LR, Watkins H. Waiting for the working draft from the human genome project: a huge achievement, but not of immediate medical use. *BMJ* 2000; **320:** 1221–22.
24 Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1·42 million single nucleotide polymorphisms. *Nature* 2001; **409:** 928–33.
25 The International HapMap Project. The International HapMap Project. *Nature* 2003; **426:** 789–96.
26 Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 2004; **5:** 89–100.
27 Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005; **366:** 941–51.
28 Cordell HJ, Clayton DG. Genetic association studies. *Lancet* 2005; **366:** 1121–31.

29 Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998; **280:** 1077–82.

30 Kruglyak L, Nickerson DA. Variation is the spice of life. *Nat Genet* 2001; **27:** 234–36.

31 Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* 2003; **33:** 518–21.

32 Goddard KA, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet* 2000; **66:** 216–34.

33 Stephens JC, Schneider JA, Tanguay DA, et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293:** 489–93.

34 Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. New goals for the US Human Genome Project: 1998–2003. *Science* 1998; **282:** 682–89.

35 Palmer LJ, Cookson WOCM. Using Single Nucleotide Polymorphisms (SNPs) as a means to understanding the pathophysiology of asthma. *Respir Res* 2001; **2:** 102–12.

36 Kruglyak L. The use of a genetic map of biallelic markers in linkage studies. *Nat Genet* 1997; **17:** 21–24.

37 Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science* 1997; **278:** 1580–81.

38 Nickerson DA, Whitehurst C, Boysen C, Charmley P, Kaiser R, Hood L. Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly informative and automatable markers for genetic linkage mapping. *Genomics* 1992; **12:** 377–87.

39 Chakravarti A. It's raining SNPs, hallelujah? *Nat Genet* 1998; **19:** 216–17.

40 McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am J Hum Genet* 1998; **63:** 241–51.

41 Kuhner MK, Beerli P, Yamato J, Felsenstein J. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 2000; **156:** 439–47.

42 Stallings RL, Ford AF, Nelson D, Torney DC, Hildebrand CE, Moyzis RK. Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. *Genomics* 1991; **10:** 807–15.

43 Brookes AJ. The essence of SNPs. *Gene* 1999; **8:** 177–86.

44 Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* 2000; **9:** 2403–08.

45 Edwards J, Bartlett JM. Mutation and polymorphism detection: a technical overview. *Methods Mol Biol* 2003; **226:** 287–94.

46 Lazarus R, Vercelli D, Palmer LJ, et al. Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease. *Immunol Rev* 2002; **190:** 9–25.

47 Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA* 1999; **96:** 15173–77.

48 Dawson E, Abecasis GR, Bumpstead S, et al. A first generation linkage disequilibrium map of human chromosome 22. *Nature* 2002; **418:** 544–48.

49 Weeks D, Lathrop G. Polygenic disease: methods for mapping complex disease traits. *Trends Genet* 1995; **11:** 513–19.

50 Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 2002; **3:** 299–309.

51 Jorde L. Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* 1995; **56:** 11–14.

52 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273:** 1516–17.

53 Toivonen HT, Onkamo P, Vasko K, et al. Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet* 2000; **67:** 133–45.

54 Templeton AR. Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome. *Ciba Found Symp* 1996; **197:** 259–77.

55 Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; **18:** 19–24.

56 Mira MT, Alcais A, Nguyen VT, et al. Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature* 2004; **427:** 636–40.

57 Thomas S, Porteous D, Visscher PM. Power of direct vs indirect haplotyping in association studies. *Genet Epidemiol* 2004; **26:** 116–24.

58 Schaid DJ. Relative efficiency of ambiguous vs directly measured haplotype frequencies. *Genet Epidemiol* 2002; **23:** 426–43.

59 Abecasis GR, Cherney SS, Cookson WOC, Cardon LR. MERLIN: multipoint engine for rapid likelihood inference. *Am J Hum Genet* 2000; **67** (suppl)**:** 327.

60 Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 2002; **70:** 157–69.

61 Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68:** 978–89.

62 Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131:** 479–91.

63 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002; **70:** 425–34.

64 Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; **12:** 921–27.

65 Morris AP, Whittaker JC, Balding DJ. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet* 2004; **74:** 945–53.

66 McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science* 2004; **304:** 581–84.

67 Ke X, Hunt S, Tapper W, et al. The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet* 2004; **13:** 577–88.

68 Phillips MS, Lawrence R, Sachidanandam R, et al. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet* 2003; **33:** 382–87.

69 Patil N, Berno AJ, Hinds DA, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294:** 1719–23.

70 Hinds DA, Stuve LL, Nilsen GB, et al. Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **307:** 1072–79.

71 Gibbs RA, Belmont JW, Hardenbol P, et al. The International HapMap Project. *Nature* 2003; **426:** 789–96.

72 Johnson GC, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29:** 233–37.

73 Goldstein DB, Ahmadi KR, Weale ME, Wood NW. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet* 2003; **19:** 615–22.

74 Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. *Trends Genet* 2003; **19:** 135–40.

75 Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF. Minimal haplotype tagging. *Proc Natl Acad Sci USA* 2003; **100:** 9900–05.

76 Ke X, Cardon LR. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* 2003; **19:** 287–88.

77 Schulze TG, Zhang K, Chen YS, Akula N, Sun F, McMahon FJ. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum Mol Genet* 2004; **13:** 335–42.

78 Patil N, Berno AJ, Hinds DA, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294:** 1719–23.

79 Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 2003; **56:** 18–31.

80 Zhang K, Calabrese P, Nordborg M, Sun F. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* 2002; **71:** 1386–94.

81 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74:** 106–20.

82 Wiuf C, Laidlaw Z, Stumpf MP. Some notes on the combinatorial properties of haplotype tagging. *Math Biosci* 2003; **185:** 205–16.

83 Weale ME, Depondt C, Macdonald SJ, et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* 2003; **73:** 551–65.

84 Ke X, Durrant C, Morris AP, et al. Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum Mol Genet* 2004; **13:** 2557–65.

85 Terwilliger JD, Goring HH. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000; **72:** 63–132.

86 Olson JM, Witte JS, Elston RC. Genetic mapping of complex traits. *Stat Med* 1999; **18:** 2961–81.

87 Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003; **4:** 587–97.

88 Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? *Nat Genet* 2000; **26:** 151–57.

89 Zhao LP, Aragaki C, Hsu L, Quiaoit F. Mapping of complex traits by single-nucleotide polymorphisms. *Am J Hum Genet* 1998; **63:** 225–40.

90 Long AD, Langley CH. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999; **9:** 720–31.

91 Lander E, Schork N. Genetic dissection of complex traits. *Science* 1994; **265:** 2037–48.

92 Wolfe KH, Li WH. Molecular evolution meets the genomics revolution. *Nat Genet* 2003; **33** (suppl)**:** 255–65.

93 Horne BD, Camp NJ. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet Epidemiol* 2004; **26:** 11–21.

94 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29:** 229–32.

95 Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002; **296:** 2225–29.

96 Crawford DC, Carlson CS, Rieder MJ, et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* 2004; **74:** 610–22.

97 Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer, 2000.

98 Molenberghs G, Williams PL, Lipsitz SR. Prediction of survival and opportunistic infections in HIV-infected patients: a comparison of imputation methods of incomplete CD4 counts. *Stat Med* 2002; **21:** 1387–408.

99 Mallinckrodt CH, Sanger TM, Dube S, et al. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol Psychiatry* 2003; **53:** 754–60.

100 Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. *Annu Rev Public Health* 2004; **25:** 99–117.

101 White IR, Moodie E, Thompson SG, Croudace T. A modelling strategy for the analysis of clinical trials with partly missing longitudinal data. *Int J Methods Psychiatr Res* 2003; **12:** 139–50.

102 Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11:** 241–47.

103 Witte JS, Elston RC, Cardon LR. On the relative sample size required for multiple comparisons. *Stat Med* 2000; **19:** 369–72.

104 Palmer LJ, Cookson WO. Using single nucleotide polymorphisms as a means to understanding the pathophysiology of asthma. *Respir Res* 2001; **2:** 102–12.

105 Rosner B. Fundamental of biostatistics. 3rd ed. Boston: PWS-Kent, 1990.

106 Lee WC. Testing for candidate gene linkage disequilibrium using a dense array of single nucleotide polymorphisms in case-parents studies. *Epidemiology* 2002; **13:** 545–51.

107 Silverman EK, Palmer LJ. Case-control association studies for the genetics of complex respiratory diseases. *Am J Respir Cell Mol Biol* 2000; **22:** 645–48.

108 Weiss ST, Silverman EK, Palmer LJ. Case-control association studies in pharmacogenetics. *Pharmacogenomics J* 2001; **1:** 157–58.

109 Palmer LJ, Cookson WOCM. Genomic approaches to understanding asthma. *Genome Res* 2000; **10:** 1280–87.

110 Feldman MW, Lewontin RC, King MC. Race: a genetic melting-pot. *Nature* 2003; **424:** 374.

111 Risch N, Burchard E, Ziv E, Tang H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* 2002; **3:** 2007.

112 Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet* 2003; **12:** 771–76.

113 Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36:** 512–17.

114 Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36:** 388–93.

115 Cavalli-Sforza LL, Menozzi P, Piazza A. History and geography of human genes. Princeton: Princeton University Press, 1994.

116 Zavattari P, Deidda E, Whalen M, et al. Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Hum Mol Genet* 2000; **9:** 2947–57.

117 Watkins WS, Zenger R, O'Brien E, et al. Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. *Am J Hum Genet* 1994; **55:** 348–55.

118 Jorde LB, Watkins WS, Carlson M, et al. Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am J Hum Genet* 1994; **54**(5)**:** 884–98.

119 Ewens W, Spielman R. The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995; **57:** 455–64.

120 Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Res* 2000; **10:** 1435–44.

121 Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001; **68:** 466–77.

122 Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001; **60:** 155–66.

123 Devlin B, Roeder K, Bacanu SA. Unbiased methods for population-based association studies. *Genet Epidemiol* 2001; **21:** 273–84.

124 Overall AD, Nichols RA. A method for distinguishing consanguinity and population substructure using multilocus genotype data. *Mol Biol Evol* 2001; **18:** 2048–56.

125 Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. *Genet Epidemiol* 2002; **22:** 78–93.

126 Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65:** 220–28.

127 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000; **67:** 170–81.

128 Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361:** 598–604.

129 Ohlstein EH, Ruffolo RR Jr, Elliott JD. Drug discovery in the next millennium. *Annu Rev Pharmacol Toxicol* 2000; **40:** 177–91.

130 Chanda SK, Caldwell JS. Fulfilling the promise: drug discovery in the post-genomic era. *Drug Discov Today* 2003; **8:** 168–74.

131 Khoury MJ, McCabe LL, McCabe ER. Population screening in the age of genomic medicine. *N Engl J Med* 2003; **348**(1)**:** 50–8.

132 Eberle MA, Kruglyak L. An analysis of strategies for discovery of single-nucleotide polymorphisms. *Genet Epidemiol* 2000; **19** (suppl 1)**:** S29–35.

133 Wright AF, Hastie ND. Complex genetic diseases: controversy over the Croesus code. *Genome Biol* 2001; **2** (8)**:**COMMENT2007.

134 Hirschhorn JN, Lindgren CM, Daly MJ, et al. Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. *Am J Hum Genet* 2001; **69:** 106–16.

135 Terwilliger JD, Weiss KM. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann Med* 2003; **35:** 532–44.

136 Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res* 1998; **8:** 1273–88.

137 Couch FJ, Weber BL. Mutations and polymorphisms in the familial early-onset breast cancer (BRCA1) gene. Breast Cancer Information Core. *Hum Mutat* 1996; **8:** 8–18.

138 Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990; **250:** 1684–89.

139 Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12–13. *Science* 1994; **265:** 2088–90.

140 Dahlman I, Eaves IA, Kosoy R, et al. Parameters for reliable results in genetic association studies in common disease. *Nat Genet* 2002; **30:** 149–50.

141 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29:** 306–09.

142 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; **33:** 177–82.

143 Tabor HK, Risch NJ, Myers RM. Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002; **3:** 391–97.

144 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ionnidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29:** 306–09.

145 Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 2001; **69:** 1357–69.

146 Mackay TF. The genetic architecture of quantitative traits. *Annu Rev Genet* 2001; **35:** 303–39.

147 Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet* 2004; **5:** 202–12.

148 Fields S. The future is function. *Nat Genet* 1997; **15:** 325–27.

149 Matsuzaki H, Loi H, Dong S, et al. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* 2004; **14:** 414–25.

150 Roses AD. Pharmacogenetics. *Hum Mol Genet* 2001; **10:** 2261–67.

151 Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 2002; **32:** 650–54.

152 Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. *Am J Hum Genet* 2004; **75:** 353–62.

153 Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* 2001; **291:** 1224–29.

154 Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999; **22:** 139–44.

155 Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002; **296:** 2225–29.

156 Judson R, Salisbury B, Schneider J, Windemuth A, Stephens JC. How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* 2002; **3:** 379–91.

157 Lowe CE, Cooper JD, Chapman JM, et al. Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun* 2004; **5:** 301–05.

158 Mallal S, Nolan D, Witt C, et al. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* 2002; **359:** 727–32.